NCI Metaphrase User Guide

Version 1.0 6/18/2001

TABLE OF CONTENTS

About This Guide

This Guide describes the NCI Metaphrase servlet Web page http://ncievs.nci.nih.gov/nci-metaphrase.html, and describes the data available in NCI Metathesaurus.

End users will find the Web servlet a useful interactive desk tool.

Software developers will find this Guide useful for the insights it provides about the intended use of the NCI Metathesaurus data. NCI Metathesaurus is a database service that is available through a Java application-programming interface. Software developers should refer to the NCI EVS Web site, http://ncicb.nci.nih.gov/core/EVS for documentation, technical advice and example code showing how to use the API.

About NCI Metathesaurus

The NCI Metathesaurus is intended to provide NCI with the complete universe of terminology relevant to NCI's operations in a single, integrated resource. Basic, translational, clinical and population-based researchers have distinct terminologies and different ways of organizing their terminologies. Therefore NCI Metathesaurus contains vocabulary from each of these domains, including a great deal of general biological, medical and epidemiological terminology. The terminologies of these four communities partially overlap, but terminological differences impede collaboration. NCI Metathesaurus "maps" terminology used in these communities to corresponding terms used in the others, thereby facilitating collaboration and sharing of information among disciplines.

The NCI Metathesaurus contains vocabularies that are part of the National Library of Medicine's Unified Medical Language System (UMLS) Metathesaurus (UMLS sources), vocabularies that NCI creates and maintains (NCI local sources) and sources that Apelon, Inc. provides as part of the NCI license for Metaphrase (Apelon sources).

The UMLS vocabularies are included to provide vocabulary that broadly covers medicine, nursing and certain other clinical disciplines and biological science.

The NCI local sources are included to address specific needs of the four communities or of the Institute. The local sources as of June 2001 are enumerated in the following table.

The Apelon sources provide convenience features such as English language definitions for much of the medical terminology or linkages to recent mentions in Medline of many concepts contained in the NCI Metathesaurus.

The NCI local sources and their intended usage are:

Vocabulary	Content	Intended Use
NCI Thesaurus	All codes, keywords and special purpose vocabulary in use within NCI	Reference terminology for internal NCI use
NCIPDQ	Expanded and re-organized PDQ	CancerLit indexing, clinical trials accrual
NCISEER	SEER terminology	Incidence reporting
CTEP	CTEP terminology	Clinical trials reporting administration, etc.
MDBCAC	Mitelman Chromosome Aberations	Cancer genome activities
Lash Tissue Terminology	NCBI tissue culture terminology	Tissue classifications in genomics and proteomics
ICDO3	Oncology classification	Cancer genome activities, incidence reporting
MedDRA	Regulatory reporting terminology	Adverse event reporting

The NCI Thesaurus versus the NCI Metathesaurus

The NCI Thesaurus is intended to become the reference terminology for cancer research. It exists as a stand-alone description logic vocabulary. In its stand-alone form, it is intended to support the Institute's software systems used for database coding and key wording, database search, data mining, text indexing and natural language understanding. The stand-alone NCI Thesaurus is provided to NCI via the Mayo Vocabulary Server, a separate service that is not part of the NCI Metaphrase service. The May 2001 version of the NCI Thesaurus contains about 18,000 concepts that are represented by about 64,000 terms. The concepts are organized into hierarchies that are up to 15 levels deep, and are also linked by about 85,000 semantic relationships.

The NCI Metathesaurus is a great deal larger than the NCI Thesaurus, containing about 850,000 concepts represented by 1,556,000 terms. There are about 4,200,000 relationships in the NCI Metathesaurus. Not all of the sophisticated semantic relationships that exist in NCI Thesaurus are available in NCI Metathesaurus. The NCI Thesaurus is included in the NCI Metathesaurus so that NCI-specific terms will be mapped to corresponding terms used in standard biomedical vocabularies.

Using the NCI Metaphrase Servlet Web Page

Go to the page http://ncievs.nci.nih.gov/nci-metaphrase.html. A servlet server running in front of the NCI Metathesaurus database generates this page. The servlet is an interactive

tool that lets you search the contents of the NCI Metathesaurus with your Web browser and navigate among vocabularies among other things.

Hint about display resolution

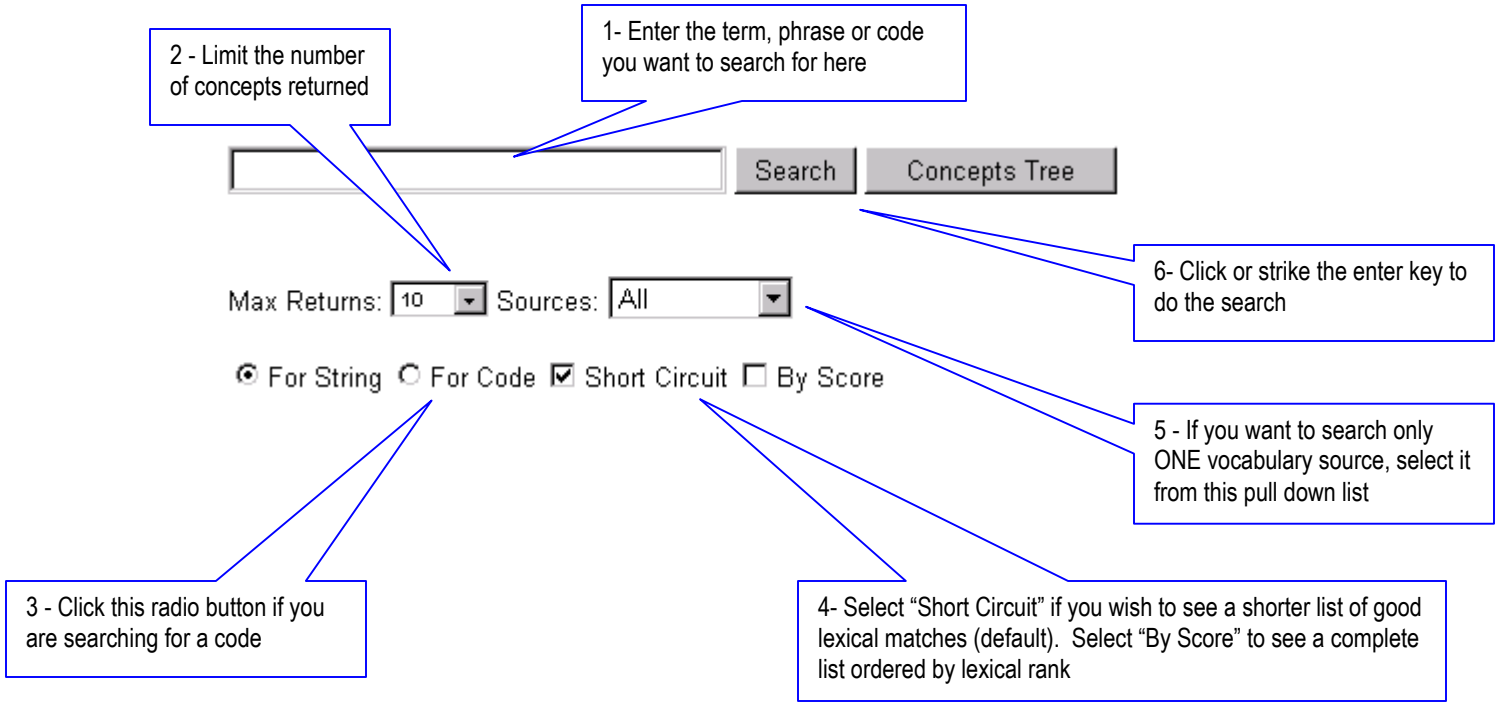
The servlet Web page looks best when you set your monitor to 1024 X 768 or greater resolution. You can display the page at lower resolution, but there will be too little room horizontally for the layout to appear without some "line wrapping". That makes the layout look messy, but all the information will still be there.

Using the NCI Metaphrase Servlet

Searching

The servlet searches using *lexical matching*, meaning that it searches for one or more authoritative terms that share "significant" *lexemes*, i.e., words or word bases, with the word or phrase you type in the search box. We say that such terms are *lexically related* to the search string (and vice versa). For example, "degenerative joint disease" is lexically related to "Joints, Knee", since they share the lexeme "joint". Of course, a closer lexical match would be "Joint disease, NOS", and an even better one would be "Disease, joint, degenerative", which is *lexically equivalent*--it contains exactly the same set of lexemes. NCI Metaphrase servlet server suggests authoritative terms for a given string by calculating a list of the most lexically related terms.

The search controls are in the upper left part of the servlet Web page. Most of the time just type the word or phrase you want to find and click "Search" or strike the enter key. The functions of the other search controls are noted in the drawing below.



Here are some tips on searching for words or phrases in NCI Metathesaurus

- Case is ignored, do not worry about capitalizing names or acronyms.
- Punctuation (except for ".", see below) is treated as whitespace (separating lexemes), and so may be replaced with a <space>, or may be omitted entirely if not in the middle of a "word".
- The words "of", "and", "with", "for", "nos", "to", "in", "by", "on", and "the" are ignored, and can be omitted.
- Word order is largely ignored, although adjacent sets of lexemes are looked up before non-adjacent sets. In general, though, don't give any thought to it.
- Rather than typing out long words, abbreviate with a period. A "." following two or more letters is treated as a *wildcard*, matching any word fragment which completes the given string of letters to form a complete lexeme. For example, typing " car. pn." will quickly return a list of matches, the first being "Pneumocystis carinii".
- You can usually omit common or redundant words, e.g., "addison" will typically return "Addison's disease" at, or near, the top of the list, and "non insulin dependent diabetes" will retrieve terms like to "Non insulin dependent diabetes mellitus" if there is no "exact" match.
- Metaphrase has spell-correction built-in, and so can usually correct simple spelling and typing errors; you can usually get the same results without going back to edit the errors in what you've already typed. Of course, in cases where there are many possible corrections, this will tend to bog down the server (since it will have to look up all of them), so you might want to correct it yourself if you don't get a speedy response.

Here are some tips to improve search speed.

- When searching for general words like “cancer” it can take a long time for the server to download all the information in the database. By limiting the search to one source vocabulary you can speed things up.
- When making heavy use of spell-correction and/or abbreviation, especially with many-word queries, setting the return limit to one or at most 10 may speed server response, sometimes dramatically.

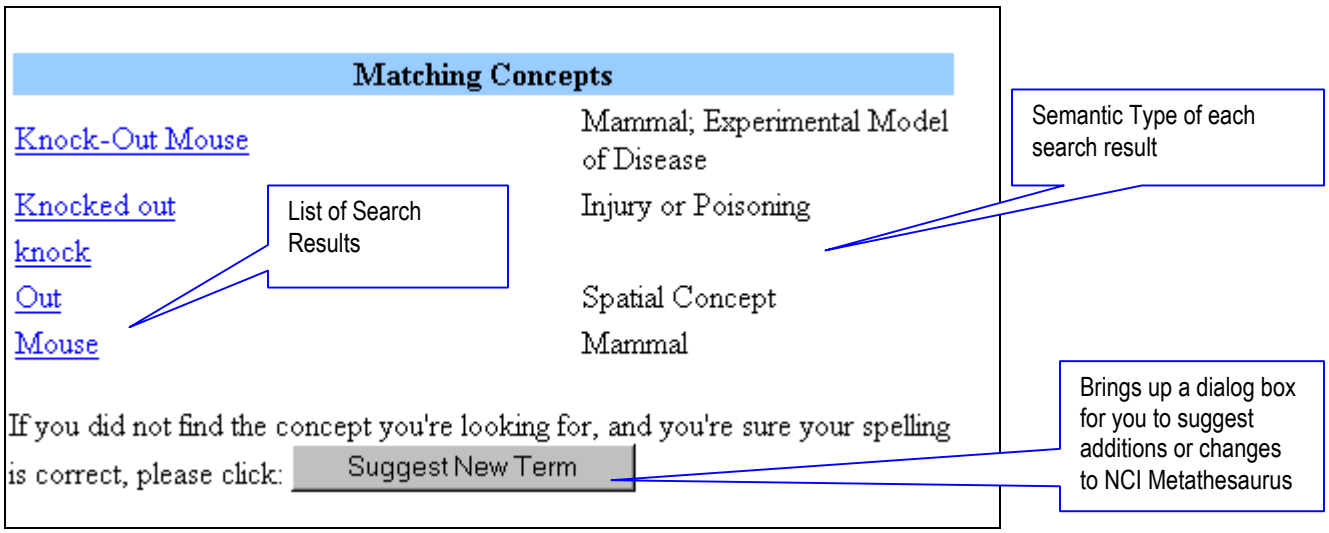
Here are some tips when searching for codes.

- Generally you will search by code when you know the code number of a term or phrase in one source vocabulary and you want to find its corresponding number in another source vocabulary.
- Enter the code you know.
- Select the “For Code” radio button.
- Select the source vocabulary from which the code came from the pull down list of “Sources”.
- The spell checking and wild card features do not work with codes. You have to enter the code correctly.

Selecting Among Search Results

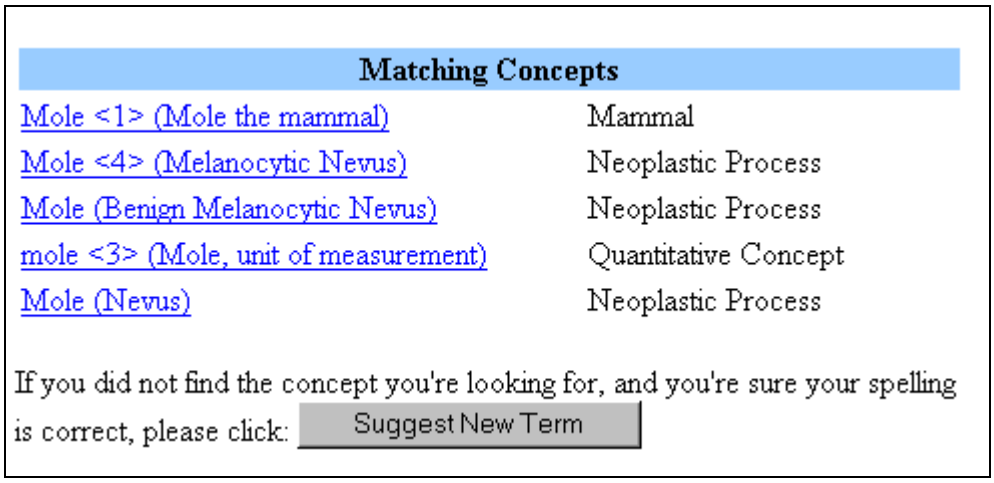
NCI Metaphrase servlet server suggests authoritative terms for a given string by calculating a list of the most lexically related terms.

This lexical matching may result in search results that are exact matches, partial matches and occasionally complete misses. If you search all sources for the phrase "knock out mouse", these are the results you will get. "Knock-out Mouse" is a complete lexically equivalent match, and as such is shown first. "Knocked out" matches two out of three lexemes, and so is shown next, even though its meaning is totally off the mark. Finally the individual words in the phrase are retrieved. This pattern of showing the most complete lexical matches first and the successively less inclusive matches in order is a characteristic of the servlet search results display.



However, perfect lexical matching does not guarantee that the results match the meaning in which you are interested[1]. Take the term "mole" for example.

[1] **Semantic** navigation means getting from a set of terms that refer to one meaning, or *concept*, to a set of terms that refer to a similar concept. All of the terms given in the example beloware *semantically related* to each other; however, the term "mole", while lexically related, are semantically unrelated. Conversely, the term "Osteoarthritis, NOS", while lexically unrelated to "Disease, joint, degenerative", is *semantically equivalent* to "Disease, joint, degenerative". That is, they both refer to the same concept. Semantic navigation in the NCI Metathesaurus is supported by an extensive, though incomplete, knowledge base of binary relationships between concepts. Concepts are usually represented for the user as a single, "NCI preferred" term.



Matching Concepts	
Mole <1> (Mole the mammal)	Mammal
Mole <4> (Melanocytic Nevus)	Neoplastic Process
Mole (Benign Melanocytic Nevus)	Neoplastic Process
mole <3> (Mole, unit of measurement)	Quantitative Concept
Mole (Nevus)	Neoplastic Process

If you did not find the concept you're looking for, and you're sure your spelling is correct, please click: Suggest New Term

If you have searched all sources in NCI Metathesaurus for the term "mole" you would have gotten the results shown above. These results will appear in the servlet Web page just below the search controls.

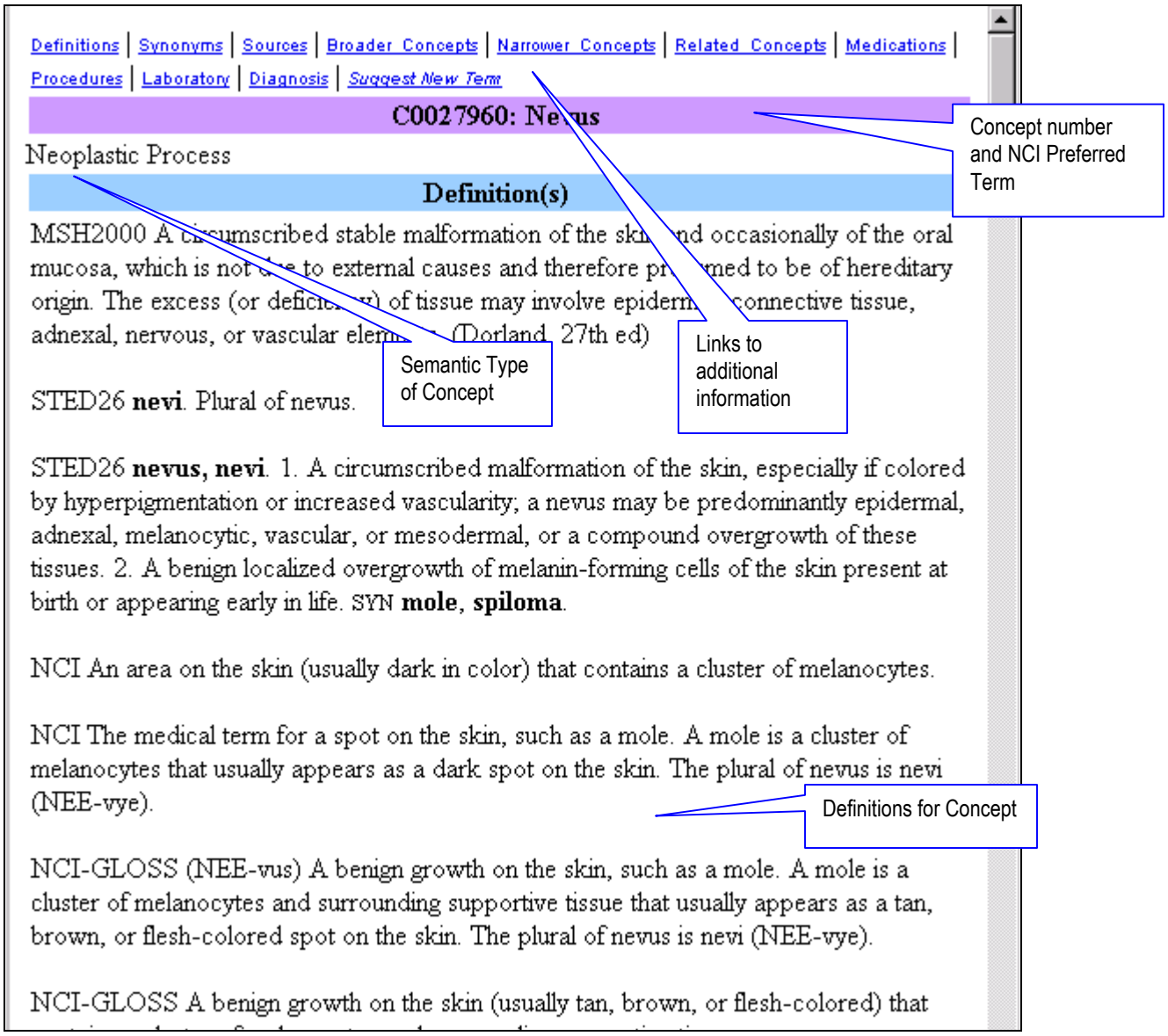
The term mole is semantically ambiguous. That is, mole can mean several things that are semantically different. In the above example, you should look at the semantic type of each search result, and click on the search result with the meaning you are interested in.

When you click on a search result, the information in the NCI Metathesaurus relating to that result is displayed in the right side of the servlet Web page. The first part of the information about "Mole (Nevus)" is shown below.

Hints on Navigation

NCI Metaphrase depends heavily on the browser's back and forward buttons (or <ALT> Left and Right Arrow if your browser buttons are not visible). The back and forward buttons affects the content of the right frame of the servlet Web page. You will find that you will follow links that appear in the right frame to view related information, and then that you will want to return to the information you had been looking at originally. Think of the information as a series of loose-leaf pages stacked up on your desk. The stack gets higher as you view new information in the right frame. The servlet Web page depends on your using the back and forward buttons to "flip" through the stack of loose-leaf pages.

You can also use the "Go To" feature of your browser to do the same thing, but experience shows that most people prefer to use the back and forward buttons.



Generally there is far more information about a concept in NCI Metathesaurus than can be displayed at one time on the right side of the servlet Web page. You can use the slider bar to the right of the information to scroll down to see the whole body of information, or you can use the links to additional information to jump directly to each sort of information.

The following Table summarizes the sort of information that NCI Metathesaurus contains. Each Row in the Table corresponds to one of the hyper-links that run across the top of the right side of the servlet Web page. Not every concept in the NCI Metathesaurus has each type of information. Tips on using each type of information are provided below.

	Type of Information	Comment
Definitions	Indicates the current concept's meaning	Sources vary in their focus, affecting definitions
Synonyms	Approximate synonyms	
Sources	Sources that contain the concept	
Broader Concepts	Parents of the concept	Not necessarily true is_a
Narrower Concepts	Children of the concept	Not necessarily true is_a
Related Concepts	Concepts with a semantic relationship to the concept	
Medications	Drugs co-occurring with concept in last three years of Medline	For concepts with a MeSH headings only
Procedures	Clinical procedures co-occurring with concept in last 3 years of Medline	For concepts with a MeSH headings only
Laboratory	Lab tests co-occurring with concept in last 3 years of Medline	For concepts with a MeSH headings only
Diagnosis	Diagnoses co-occurring with concept in last 3 years of Medline	For concepts with a MeSH headings only
Suggest New Term	Calls up a dialog box. You enter your suggestions, comments, etc. Our editors will follow up and contact you.	Enables you to request changes or additions to NCI Metathesaurus

Tips on using the information in NCI Metathesaurus

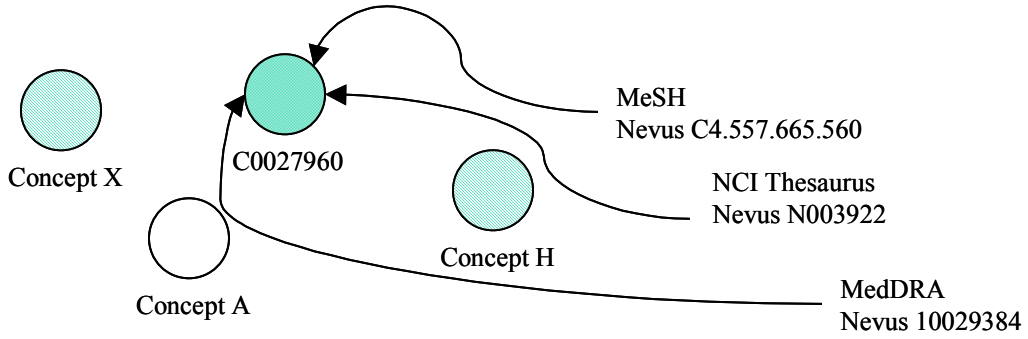
NCI Thesaurus and other NCI Local Sources

The version of the NCI Thesaurus in the NCI Metathesaurus has sub-sources. This is unique among source vocabularies in NCI Metathesaurus. These sub-sources represent information that is specific to only a part of the Institute, for example RAEB. The major affect of these sub-sources is that they provide a way to identify definitions that are "owned" by RAEB or some other part of the Institute.

Concepts versus Terms and their Role in Vocabulary Mapping

The mechanism used in NCI Metathesaurus to map between a term in one source and corresponding terms in other sources is to assign each term to a concept. All the terms and phrases in NCI Metathesaurus are linked to one or more concepts. A concept in NCI Metathesaurus is simply a unique reference number. In NCI Metathesaurus concepts become meaningful because of the information associated with them. At a minimum, all concepts have a semantic type and an associated "term" in addition to a unique concept number.

In the example in the following figure, the concept C0027960 has the semantic type "Neoplastic Process", denoted by the solid green fill color. "Atoms" or terms from three



sources are mapped to it. (An "atom" is simply a term in a specific source). This means that the NCI EVS editors or the NLM UMLS editors have determined that the MeSH atom Nevus, the NCI Thesaurus atom Nevus and the MedDRA atom Nevus all have the same meaning. Within a source, each term or "atom" has an "atom ID number". In the example, the Mesh atom ID is C4.557.665.560.

The Concepts X and H in the above drawing have the same semantic type, which is not the same as either Concept A or Concept C0027960. For example, Concept A might have the semantic type "Mammal" while the other two might be semantic type "Quantitative Process".

In summary, terms within sources have atom ID numbers that uniquely identify the term within that source. Atoms are associated to concept numbers, which are unique across the entire NCI Metathesaurus. Finally, all concepts are of at least one semantic type. It is these formal properties of the NCI Metathesaurus database, along with the expertise of the NCI EVS editors, which enable you to translate from one source to another.

The Four Types of Concept Number and Coding or Keywording

One major difference between NLM UMLS Metathesaurus and the NCI Metathesaurus is that in UMLS Metathesaurus all concepts have concept numbers of the form C#######. In NCI Metathesaurus there are sources that do not appear in UMLS, such as Stedman's Medical Dictionary and the NCI local sources. Therefore in NCI Metathesaurus you will see concept numbers of the form CL#######, CS####### and CF####### as well as C#######. All four concept number formats function the same, but the CL lets you know that the concept occurs only in a NCI local source, CS that it occurs only in Stedman's Medical Dictionary, and CF that it occurs only in First Data Bank's National Drug Data Codes (NDDF).

If you are using NCI Metathesaurus to obtain concept numbers that you use in coding or key wording data, **we recommend you use only C###### and CL####### numbers as codes or keywords**, as the CS and CF numbers may disappear in the future if licensing for the Dictionary and NDDF is not available in future years.

Definitions

Many concepts in NCI Metathesaurus have definitions from multiple sources. In general they will differ in wording or emphasis but not in meaning. If you encounter a concept with definitions that frankly differ in meaning it is an error on our part. Please call it to our attention.

Regarding the NCI local sources, most NCI Thesaurus atoms have a definition that is labeled "NCI". In addition, the term may have other definitions labeled, for example, "RAEB" or "DBS". Most of these branch- and division-specific definitions will be disappearing during FY01. They are being replaced with consensus NCI definitions.

All NCI atoms will eventually have a consensus definition, which will be labeled "NCI". These "NCI" definitions are intended to be scientifically correct, precise definitions. Some NCI atoms will have a second definition labeled "NCI-GLOSS". NCI-GLOSS is short for NCI Glossary. These definitions are generally correct definitions requiring a lower reading level. They are intended suitable for the lay reader of NCI Web pages and so on. A few NCI atoms will also have "RAEB" definitions. These may not be true definitions but rather a sort of "scope note" intended to assist coders in using the RAEB coding terms correctly.

Synonyms

Synonyms are displayed in the right of the servlet Web page below the definitions. The synonymy displayed for all sources should be interpreted as being approximately correct. In many cases there are fine distinctions to be made between nominal synonyms, and in some there are synonyms that are arguably incorrect. This lack of precision is unavoidable given the imprecise nature of English and the limitations of our semantic tools for mapping among atoms. Nonetheless the synonyms are valuable because the synonyms enable one to search using any one of many terms and to generally find the right underlying concept. (See Mapping, below.)

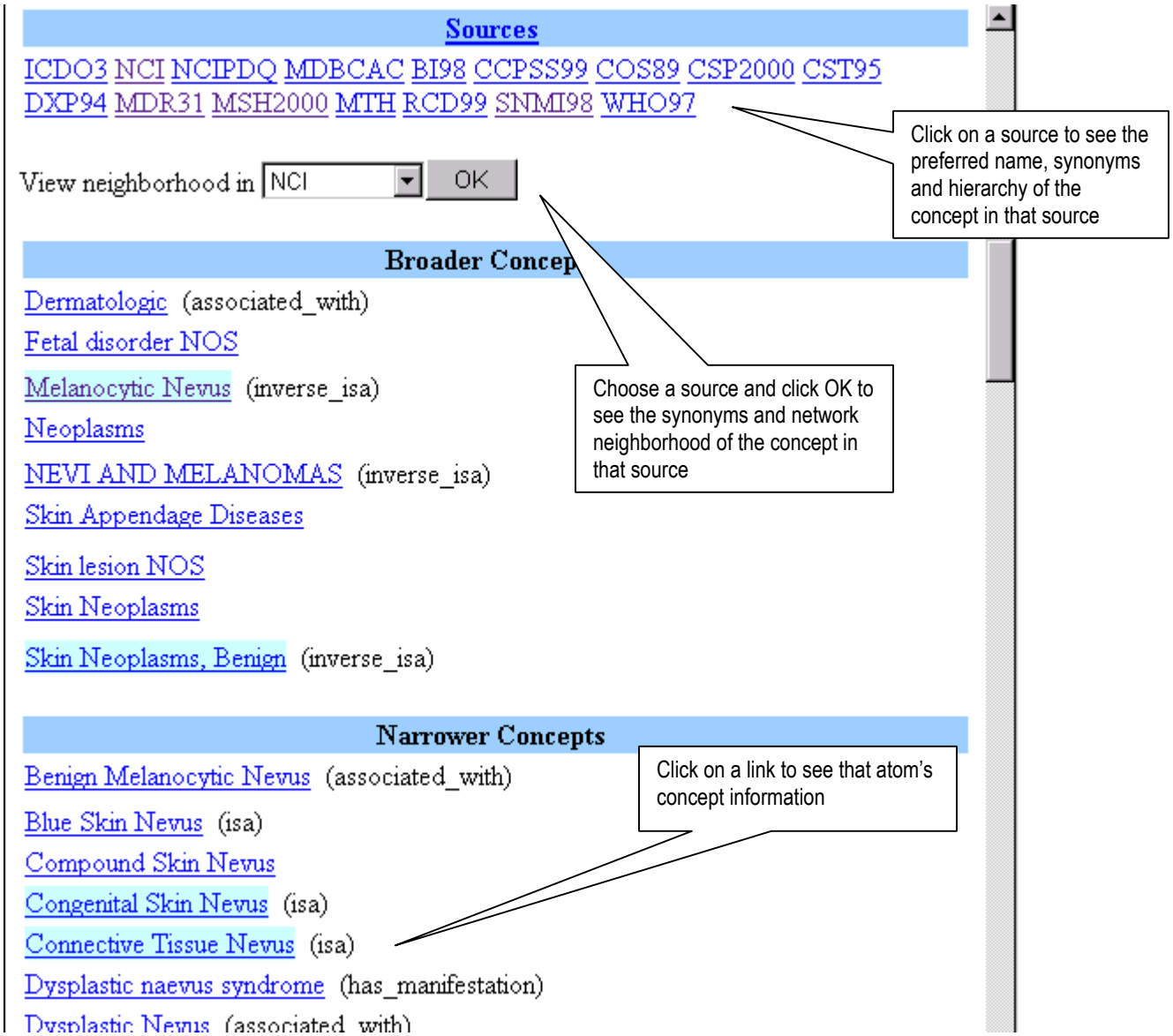
Sources

Below the list of synonyms is a list of the sources that contain one or more atoms mapped to the concept for which you searched. These sources are denoted by their acronyms. (You can check the meaning of a source acronym by scrolling down the top-left frame of the servlet Web page. To do so, open the servlet Web page again in a new browser window.)

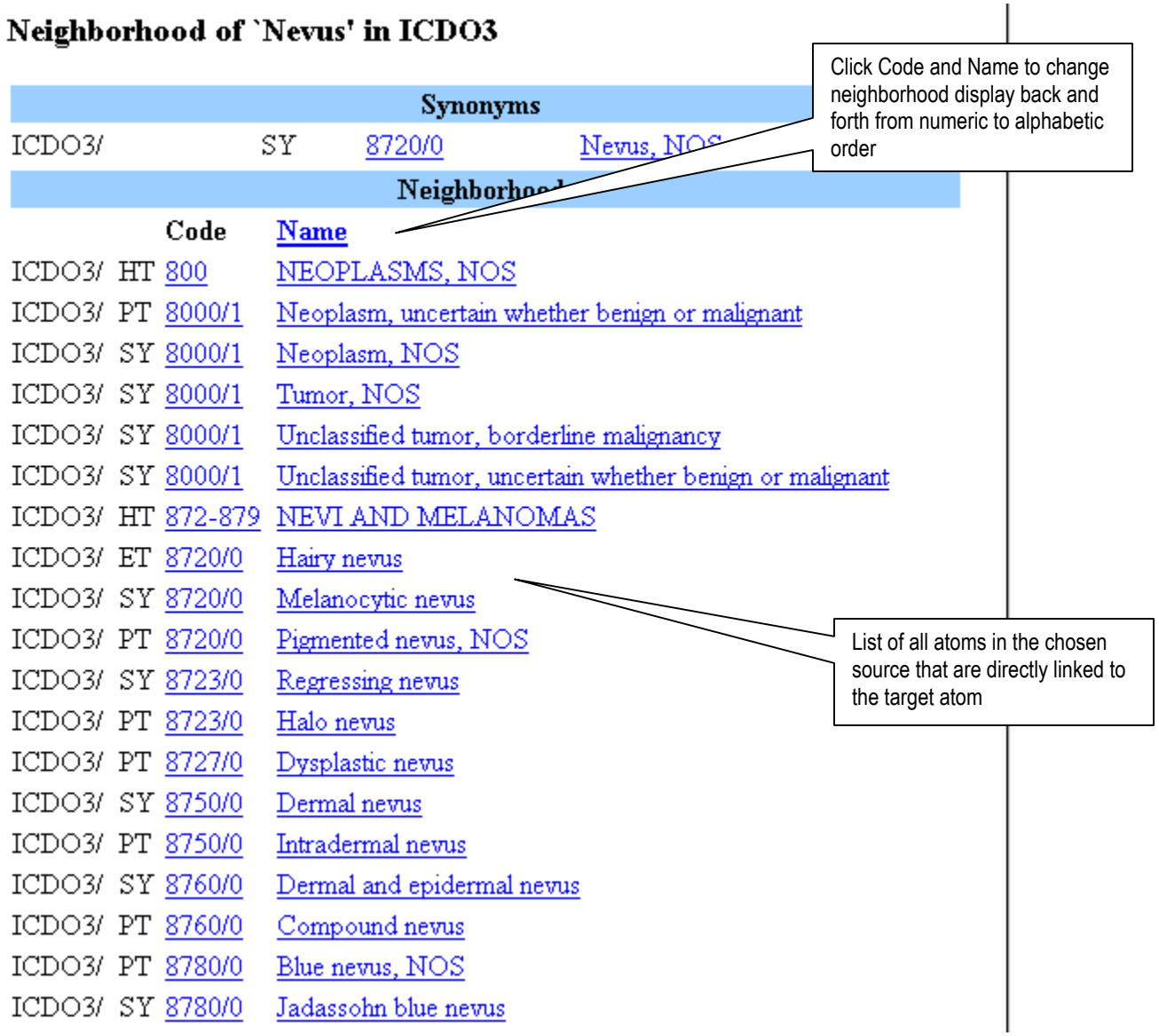
The source acronyms are all links. If you click on one, the right frame of the servlet Web page will display the preferred term, synonyms, and acronyms and their atom ID numbers, and in some sources, part of the hierarchy from the source you click on[2]. (Use the back button or ALT_Left_Arrow to restore the display of sources.

[2] See **Viewing the NC Thesaurus Hierarchy** at the end of this document.

Below the list of sources you will see "View Neighborhood in" followed by a pull down list of sources and an "OK" button. If you select a source in the pull down list and click OK, you will see the same list of synonyms as you would see had you clicked on the source's hyperlink, but below the synonym list you would see a (usually long) list of "semantically related" atoms from the source. The notion of neighborhood is easy to define, it is the list of all atoms within the source you pick from the pull down that are a distance of one from link from the atom you are displaying when you click "OK". However since in NCI Metaphrase the meaning of the links differs depending on the source, the neighborhood displayed may be perfectly rational and easy to understand in one source and puzzling in another.



The NCI Thesaurus neighborhoods typically consist of the NCI atoms associated with concepts that are immediately broader and narrower than the concept you were displaying when you click "OK". The neighborhoods from sources such as ICDO3, which have



atom ID numbers that specify hierarchy location, often provide the more interesting neighborhoods.

In the above example, ICDO3 was chosen as the source, and nevus as the atom. By default the atoms in the neighborhood are displayed in alphabetical order. However you can display them ordered either by the numerical order of their atom ID numbers, or by the alphabetical order of their names. In the above example they are displayed by numeric order. By clicking on the blue underlined "Name" link one could display them in alphabetic order. If you did the bold "Code" would turn into a link, and by clicking it, one could return the display to numeric order.

Broader Concepts

The list of broader concepts shows the next broader concept from every source that contains the current concept. Put another way, if a source contains an atom that is linked to the current concept, and if the source is a hierarchical source, there will most likely be a atom in the source that the source authors consider more general than the atom linked to the current concept. The broader concepts list shows each of these broader atoms.

Narrower Concepts

The list of narrower concepts shows the next more specific concept from every source that contains the current concept. Again, if a source contains an atom that is linked to the current concept, and if the source is a hierarchical source, there will most likely be a atom in the source that the source authors consider more specific than the atom linked to the current concept. The narrower concepts list shows each of these narrower atoms.

Related Concepts

The list of related concepts shows the semantically related concepts from every source that contains the current concept. Again, if a source contains an atom that is linked to the current concept, and if the source has semantic relationships, there may be a atom in the source that the source authors consider related in some way to the atom linked to the current concept. The narrower concepts list shows each of these semantically related atoms.

If broader and narrower relationships organize concepts vertically, semantic relations can be thought of as relating them horizontally. Semantic relationships in some sources are very sophisticated and specific, representing ontologic information such as the fact that a specific bacterium is the etiologic agent that causes a specific disease. In other sources there are only some primitive indication that some sort of undefined relationship exists among two or more concepts.

The NCI Metathesaurus cannot differentiate among these relationship, but lumps the sophisticated precise relationships from sources such as the NCI Thesaurus with the vague ones. The way to think of the related concepts in NCI Metathesaurus, is that in one or more source, some sort of horizontal relationship is stated between the current concept and the concepts whose atoms appear in the list of related concepts.

Navigation in the Broader, Narrower and Related Concepts Lists

All the atoms listed in the broader, narrower and related lists are hyperlinks. If you click on one, the contents of the right frame of the servlet Web page changes to display information about the concept to the atom on which you clicked. Use the page back commands of your Web browser to restore the information that was displayed before you clicked.

Some of the atoms in the broader, narrower, and related lists have a notation appended to them that describes the specific type of broader relationship (for example, inverse is_a, is_associated_with), narrower relationship (is_a, part_of) or related relationship (is_mapped_from). We display this information when it is available in the source. It is intended to help you determine if you want to follow the link to view the information on the concept pointed to by the link.

Some of the atoms in the three lists are highlighted in blue. These are atoms from the NCI local sources.

This servlet supports *lexical* and *semantic* navigation of the NCI Metathesaurus terminology database.

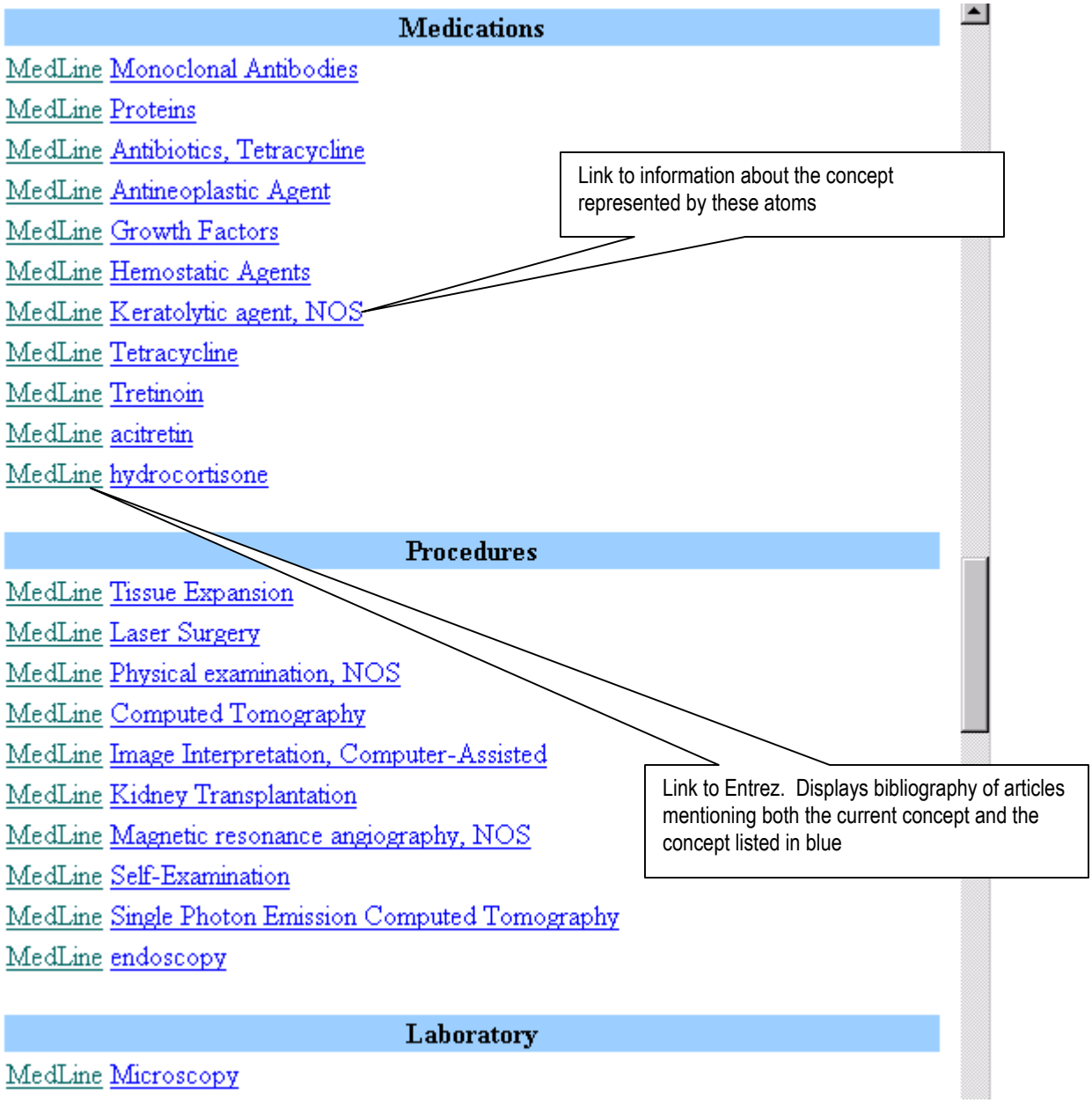
Medications, Procedures, Laboratory, and Diagnosis Lists

All concepts in NC Metathesaurus that have atoms in MeSH that are MeSH headings will have lists of links in each of the medications, procedures, laboratory and diagnosis sections at the bottom of the right side of the servlet Web page.

We index the most recent three years of Medline to create a list of citations. In each citation, the current concept co-occurred with one or more concepts contained in the MeSH that denote a medication, a diagnostic procedure, a clinical laboratory test, or a clinical diagnosis.

There are blue hyperlinks and green hyperlinks displayed. The green links are links to Medline. If you follow the green link, a new browser window opens, the National Center for Biotechnology Information Entrez site is contacted, and a query is executed that causes a list of all citations in Medline to be displayed that mention both the current concept and the blue linked concept to the right. If you follow the blue link, the right frame of the servlet Web page will display the concept information for the atom you click on.

This feature of the servlet Web page is intended to assist those interested in finding a quick bibliography of clinically relevant articles, or as a way to find out more about the meaning of concepts and how they fit into the clinical enterprise.



Viewing the NCI Thesaurus Hierarchy

Some sources have atom numbers that specify the location of each atom in the source's hierarchy. Examples include PDF 2000, MeSH and Snomed Version 3. When you display the current concept in one of these sources[3], you will see the fragment of the source hierarchy from the current concept back to the top, or "root" of the source's hierarchy tree. This is useful in seeing where a concept fits in the hierarchy of the specific source.

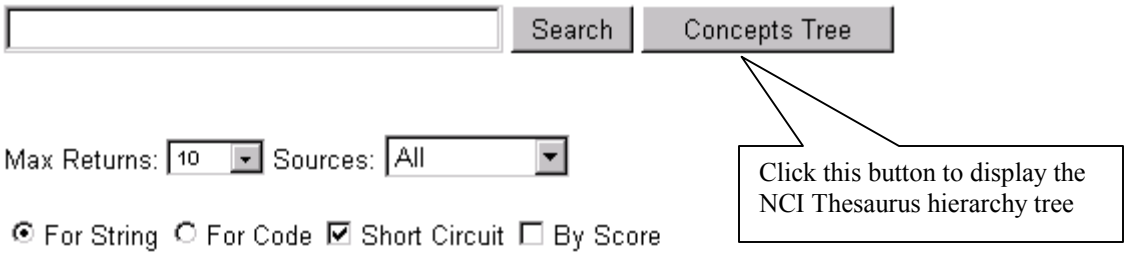
[3] See **Sources,** above

There are at least two issues that you should be aware of, one relating to the limitation of the NCI Metaphrase software, and the other related to a change in the philosophy and practice of vocabulary construction.

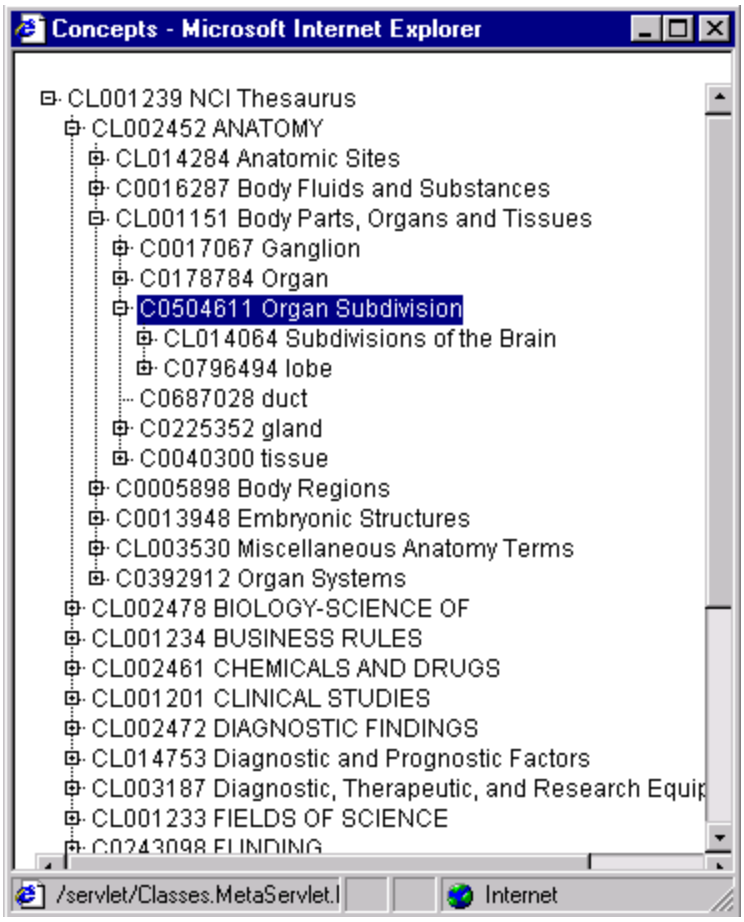
When you display the current concept in a source, you will see a hierarchy fragment only for sources that use atom numbers that specify hierarchical location. Furthermore, NCI Metaphrase has no means of letting you browse the entire hierarchy of any source. At best you will see the fragment from the current concept to the root.

The sources that use hierarchical numbering have found that it causes them serious problems. If a lot of new information is discovered that results in a large number of new atoms being added to an already heavily populated part of their hierarchy, they run out of numbers to assign to the new atoms. The labor expense involved in renumbering is too great to be tolerable. Therefore most major vocabulary publishers are moving to use of numbers that do not specify tree position.

The NCI Thesaurus and our other local sources use numbers that are unique, but do not specify hierarchy location. Nonetheless we want to provide a means for NCI staff to browse the NCI Thesaurus from within the NCI Metaphrase Web page. We have provided a Java applet to serve this function.



A new window is created for the NCI thesaurus tree display. The NCI tree display allows you to expand the hierarchy by clicking on the plus icons. If you click on an atom in the hierarchy, the information for the concept represented by that atom will appear in the right frame of the servlet Web page.



For more information, visit http://www.apelon.com or http://www.lexical.com/, or email info@lexical.com.